DOCUMENT RESUME

ED 190 622 TM 800 416

AUTHOR Nassif, Paula M.; And Others
TITLE Further Work Developing Parallel Tests by P-Value Item Substitution.
PUB DATE Apr 80
NOTE 18p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Boston, MA, April 8-10, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary Secondary Education; *Equated Scores; Hypothesis Testing; Minimum Competency Testing; *Quantitative Tests; *Raw Scores; *Reading Tests; *Scaling; Testing Programs
IDENTIFIERS *Linear Equating Method; New Jersey Minimum Basic Skills Programs; *Parallel Forms Reliability

ABSTRACT

The statistical parallelism of a pair of tests equated by the P-value method was investigated, as well as the relationship between their raw scores and scores transformed by linear equating. Data were obtained from eight sections of the 1980 New Jersey Minimum Basic Skills Program: specifically, the reading and mathematics tests for each of grades 3, 6, 9, and 11. Twenty-five equating test items were selected intact, and 25 were selected by p-value item substitution, as the parallel forms. Both the Wilks test and linear score equating supported the parallelism hypothesis for the two subject areas and four grade levels, except for the grade 12 mathematics test. Thus, the p-value method would benefit testing programs, where two or more parallel forms are administered annually for practical and statistical reasons. First, p-value requires a smaller sample size than linear or equipercentile score equating. Second, the technique of adding pilot items to existing forms can be used to generate, during an actual administration, the item statistics (p-value and point biserials) needed for item-substitution. (CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED190622

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

FURTHER WORK DEVELOPING PARALLEL TESTS BY P-VALUE ITEM SUBSTITUTION

Paula M. Nassif
Paul D. Pinsky
Sherry A. Rubinstein

National Evaluation Systems, Inc.
Amherst, Massachusetts 01002

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Paula M. Nassif

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the Annual Meeting of the National Council on Measurement in Education, Boston, Massachusetts, April 1980.

TM 800416

FURTHER WORK DEVELOPING PARALLEL TESTS
BY P-VALUE ITEM SUBSTITUTION

Background

The past decade has seen a range of test development activities in the political and policy environments of state departments of education. Statewide testing, of either the sampling or census variety, has been conducted for reasons of public educational accountability, general curricular or program evaluation, and individual student assessment. State departments undertaking the development of customized instruments have typically commissioned development of a single form of a test to assess a particular set of domains or objectives. Relatively little need was seen for the development of alternative or parallel forms since, even in ongoing or cyclical programs, each test administration involved either a new sample of students, a new population of target grade students, or testing in an entirely new subject area.

The advent of minimum competency testing has altered the complexion of many state-level test development efforts. Minimum competency testing has brought with it two important changes: (1) wider application of repeated testing of the same student to certify the efficacy of remediation or simply to provide more than one opportunity to achieve a passing score, and (2) increased politicization of the tests and testing process because, more so than in the past, test results affect very directly the lives of individual students. Repeated testing of the same student raises questions about the extent to which recollection of specific test items from a prior administration can invalidate or contaminate a student's test scores on a subsequent administration. The highly political nature of some testing programs raises questions about unauthorized duplication and dissemination of test items and increases the likelihood of legal challenges or truth-in-testing legislation that would require baring tests to public inspection.

One response to these issues, and perhaps the most prevalent, has been an increased emphasis on development of parallel forms of state tests. This response is understandable for three reasons. First, the availability of parallel forms reduces the problem of test security between administrations. Second, it answers political pressure to release the test after administration for use in diagnosis of student weaknesses and tailoring of remedial services. Third, it ensures that an individual student may be retested on the same skills with different test items, minimizing the effects on performance of the prior administration.

The increasing need for alternate forms of tests in states across the country has redoubled interest among researchers, educators, and policymakers in how best to ensure that the score or pass-fail decision for a given student not depend on "which form" the student took or "when" the student participated. The statistical problem of test form equivalence takes two primary forms. The first is maximizing the likelihood that a student would receive the <u>same score</u> on two different forms of a test. The second is a more simplified task of minimizing errors of classification--that is, maximizing the likelihood that a student will receive the same classification (pass or fail), although not necessarily the same score, on two alternate forms. The former is most appropriate when the purpose of testing and the prescribed use of test results is to analyze a student's <u>level</u> of functioning and compare it from administration to administration. The latter is most typical of minimum competency testing programs that are directed primarily at determining a student's status simply with respect to a cut-score.

While there may be some debate as to which of these statistical problems is most relevant in a particular measurement situation, the debate rests on the common assumption that parallel forms must conform statistically in order for one to draw valid inferences from the results of retesting a student. There remains, however, the central question of how that statistical conformity should be achieved.

In practical terms, there are two approaches to accomplishing statistical equivalence of alternate forms. One is to "equate tests" by selecting items with equivalent psychometric characteristics; for example, the Rasch model (Wright, 1977) and the p-value method (Nassif, Rubinstein, & Pinsky, 1979). The other is to "equate scores" by paying relatively less attention to the psychometric characteristics of individual items (except in the normal course of screening for psychometric adequacy) and solving the statistical problem by scaling the test (or subtest) scores produced; for example, the linear and equipercentile methods (Angoff, 1968.)

It should be noted that regardless of which statistical approach is used, the issue of content equivalence can be addressed in the same manner: where minimum competency tests are objective- or domain-referenced, for example, the same numbers and types of items can be selected to measure each construct on each alternate form. Such "parallel blueprinting" techniques can be used to establish content equivalence in tandem with either score or test equation methods. The focus of this paper is on issues of statistical, not content, equivalence of test forms.

Purpose

By comparison with the "score equating" methods which traditionally have been used in developing alternate forms of standardized tests, "test equating" methods are relatively recent in the testing field. Consequently, the Rasch and p-value methods have come under careful scrutiny as they have become more widely used. The examination has centered on the question of whether, in comparison to the more traditional methods, these approaches achieve the same results or demonstrate the same statistical adequacy. This paper addresses these questions specifically with regard to p-value test equating and linear score equating.

The two primary purposes of the study are: (1) to determine whether the statistical properties of a pair of tests developed (equated) by the p-value method conform sufficiently to permit the conclusion that the tests are statistically parallel, and (2) to examine the relationship between the raw scores produced by those tests and the transformed scores that would result from linear equating of the scores.

More generally, the purpose of the paper is to provide further information about the technical adequacy of the p-value method in order to more fully consider the feasibility or desirability of using the method in light of its other virtues in applied settings.

Design

The study described in this paper was based on a set of data obtained in the course of a field test of new test items for the 1980 New Jersey Minimum Basic Skills Testing Program. Included in the program is a set of eight tests--reading and mathematics tests for each of grades 3, 6, 9, and 11. Each of the four mathematics tests contains approximately 100 items distributed across approximately 50 specifications and four clusters. Each of the four reading tests contains approximately 100 items distributed across approximately 20 specifications and three clusters.

The first set of eight tests having been developed for the 1978 testing program, the linear score-equating method had been used to ensure the equivalence of the 1978 and 1979 versions of the tests. The same method was being used to score-equate the 1980 forms: for each of the eight 1980 tests, about 125 new items were developed and field-tested in tandem with a 25-item "equating test" composed of items from the respective 1979 version. Each "equating test" was comprised of a representative subsample of the original full-length test.

In the Fall of 1979, two stratified random samples of approximately 500 New Jersey students in each of grades 4, 7, 10, and 12 was selected to participate in the field test of the items designed for spring testing of grades 3, 6, 9, and 11, respectively. To each sample of 500 students was administered a grade-appropriate 150-item (125 new items plus 25 "equating items") field test form. It was from this pool of data that data were drawn for the present study.

For purposes of the study, two research "mini-forms" for each of the eight tests were created. In each case, the 25 "equating" test items were selected intact to represent Mini-Form A and a set of 25 items from the respective 1980 item pool was selected by the p-value item substitution method to comprise Mini-Form B. For each item on a given Mini-Form A, an item with an equivalent p-value (or the item with the closest p-value) was selected for the corresponding Mini-Form B. (It should be recalled that the p-values for both the equating items and for the items in the 1980 pool reflected the performance of the same students--those who participated in the field test.) For the reading tests, each item on Mini-Form B assessed the same specification as its parallel item on Mini-Form A. For the mathematics tests, it was occasionally necessary to select parallel items for Form B from different specifications, although still within the same cluster.

The result of the test construction task was eight pairs of Mini-Forms A and B: a total of 16 separate 25-item research forms. Student responses to these 400 items were then extracted from the larger data set and each of the 16 mini-forms was scored. The distribution of scores was highly skewed for all 16 tests. On the reading tests the mode was approximately 24 and the means about 22, while on the mathematics tests the mode was approximately 22 and the mean about 20 on the 25-item tests. The standard error of measurement ranged from 1.34 to 1.52 for the reading tests, and from 1.64 to 1.82 for the mathematics tests.

Analysis

Statistical test of parallelism. The first step of the analysis was to select an appropriate test of statistical parallelism for Mini-Forms A and B of each pair of tests. Taking the definition of parallelism as the condition under which the probability distribution of score results is the same for the same examinee on both test forms, a statistic developed by Wilks (Gulliksen, 1950) was adopted. The Wilks statistic was applied because it provides for a stricter, more conservative test of the parallelism of the forms than a test of the equality of their means (using a t-test) and the equality of their variances (using an F-test). As may be seen by the formula below, the Wilks statistic incorporates the correlation between the two forms; it therefore negates the possibility that two test forms with equal means and equal variances, but having a strongly negative correlative relationship, could meet a test of parallelism.

The Wilks statistic, which assumes a normal distribution and is approximately distributed as chi-square with 2 degrees of freedom, is:

$$-N \log_e L_{mvc}$$

where

$$L_{mvc} = \frac{S_A^2 S_B^2 - C_{AB}^2}{S^2 (1 + R) (S^2 (1 - R) + V)}$$

and

$$S^2 = \frac{S_A^2 + S_B^2}{2}$$

$$R = \frac{C_{AB}}{S^2}$$

$$V = (M_A - M)^2 + (M_B - M)^2$$

$$M = \frac{M_A + M_B}{2}$$

M_A, M_B = mean of Test A, mean of Test B

S_A^2, S_B^2 = variance of Test A, variance of Test B

C_{AB} = covariance of Test A and Test B

Linear equating. The second analysis of the study was to linearly equate each pair of research forms. For each of the eight pairs, Mini-Form B was linearly equated to Mini-Form A, thus expressing the raw scores on B in the scale of A. The equation shown below (Angoff, 1971), assuming equal first and second moments, was used:

$$\frac{A - M_A}{S_A} = \frac{B - M_B}{S_B}$$

and thus

$$A = \frac{S_A (B - M_B)}{S_B} + M_A$$

While all of the pairs were equated as above, this paper presents, for illustrative purposes, the linear equating results for only two pairs, the pairs that represented the most extreme values of the Wilks statistic: the "most parallel" and the "least parallel."

Results

Results of the Wilks test. Displayed in Table 1 are the results of the Wilks analysis. Given the null hypothesis that there is no significant difference between Mini-Form A and Mini-Form B of each pair, a critical value of 5.99 would be required to reject the null hypothesis with 95% confidence. Table 1 shows that, at the 95% level, the null hypothesis is rejected for only one of the eight pairs of tests (grade 12 mathematics); for the other seven pairs of tests there was a failure to reject the null hypothesis. At the 99% confidence level (critical value 9.21), wherein it is more difficult to reject the null hypothesis, the null hypothesis was not rejected in any case. None of the pairs of tests were "different enough" to permit rejection of the null hypothesis at this level of confidence.

Of the eight pairs of tests, the "worst case" (the pair with the highest value of the Wilks statistic) was the grade 12 mathematics test. The "best case" (the one with the smallest Wilks value) was the grade 7 reading test. The results of linear equating of Forms A and B for each of these pairs is presented below; the presentation assists in the interpretation of the Wilks statistic insofar as one may observe, for each Wilks value, the corresponding differential between the raw and equated scores.

Results of linear equating. The linear equating results are presented in Table 2 for the grade 7 reading test and in Table 3 for the grade 12 mathematics test. For each possible raw score (0 through 25) on a given Mini-Form B, the tables show the score that "would have been achieved" on Form A--that is, the raw scores of Form B expressed in the scale of A.

In Table 2, the linear equating shows very close score equating over the entire range, while in Table 3 the scores at the lower end of the distribution differ by up to 1 point. It should be noted that all differences of scores are well within the standard error of measurement of the tests.

TABLE 1

Results of Wilks Statistical Test of Parallelism

	Mean		Variance		R	C_{AB}	Wilks $-N \log_e L_{mvc}$
	Form A	Form B	Form A	Form B			
Reading: Grade 4	22.48	22.45	8.37	9.39	.83	7.35	4.89
Math: Grade 4	19.81	19.73	23.85	23.34	.92	21.62	1.12
Reading: Grade 7	21.48	21.50	9.61	9.42	.77	7.28	0.18
Math: Grade 7	19.42	19.46	20.29	18.80	.87	17.00	3.16
Reading: Grade 10	22.00	21.99	13.00	14.12	.86	11.61	2.97
Math: Grade 10	19.33	19.35	17.30	18.61	.88	15.89	3.97
Reading: Grade 12	21.00	21.15	17.24	16.51	.86	14.58	0.92
Math: Grade 12	21.16	21.13	17.60	15.77	.88	14.82	6.87

NOTE: The 5% critical value of the Wilks statistic (chi-square with 2 degrees of freedom) is 5.99, while the 1% critical value is 9.21.

TABLE 2

Linear Equating for Grade 7 Reading Tests

Raw Score on Form B	Equated Score in Scale of Form A
25	25.02
24	24.01
23	23.00
22	21.99
21	20.98
20	19.97
19	18.96
18	17.95
17	16.94
16	15.93
15	14.92
14	13.91
13	12.90
12	11.89
11	10.88
10	9.87
9	8.86
8	7.85
7	6.84
6	5.83
5	4.82
4	3.81
3	2.80
2	1.79
1	0.78
0	- 0.23

TABLE 3

Linear Equating for Grade 12 Mathematics Tests

Raw Score on Form B	Equated Score in Scale of Form A
25	25.25
24	24.19
23	23.13
22	22.08
21	21.01
20	19.97
19	18.91
18	17.85
17	16.79
16	15.78
15	14.68
14	13.62
13	12.57
12	11.51
11	10.45
10	9.40
9	8.34
8	7.28
7	6.22
6	5.17
5	4.11
4	3.05
3	2.00
2	0.94
1	- 0.12
0	- 1.17

Discussion

The results of the Wilks test warrant discussion. Typically, a researcher setting a null hypothesis that A "is equal to" B is attempting to reject the null hypothesis; the 99% confidence level in the typical case provides a stricter, more rigorous criterion than the 95% confidence level, making it more difficult to obtain the desired experimental outcome. This study entails the unusual situation in which the desired outcome is to avoid rejecting the null hypothesis. One peculiarity of this case is that increasing the confidence level (for example, from 95% to 99%) works to the favor of, rather than against, the researcher. The greater the difference (between Forms A and B) "allowed by" the level of confidence, the greater will be the chance that Forms A and B will be "accepted as parallel." In the face of this anomaly, the 95% and 99% confidence levels were a priori adopted on the basis of convention.

Notwithstanding the weaknesses of hypothesis testing as a method in this case, the results of the Wilks test are supportive of the outcomes of the p-value approach. With only one exception, "parallelism" was shown for tests in two difference subject areas and at four different grade levels. From the breadth of these cases, one may have sufficient confidence that using the p-value method will yield a test that fulfills the classical definition of parallelism to the extent that it can be assessed.

The linear equating analysis was directed as discovering, for a given score on a test developed (a priori equated) by the p-value method, what the corresponding a posteriori equated score would be. With few exceptions, the answer for all eight pairs of tests was that each raw score and corresponding related score are virtually identical when scores are rounded to the nearest whole score. The exception to this pattern was in the scores at the tail of the distribution of the grade 12 mathematics test (raw scores 0 through 7 on a 25-item test, with a marked skew). Given that this test represented the

"worst case" of the Wilks statistic there was remarkably close score equating across the rest of the range (raw scores 8-25).

It should be noted, however, that in those testing environments in which the primary statistical problem is consistency of classification (pass-fail decisions) rather than specific score correspondence, even the grade 12 mathematics test might display acceptable characteristics. In the New Jersey testing program, the cut-scores are 75% in reading and 65% in mathematics. These translate to 18 of 25 correct on the reading mini-tests and 16 of 25 on the mathematics mini-tests. If these cut-scores were adopted for the mini-forms, the p-value method would have provided consistent classification when the data were rounded to the nearest score.

Conclusion

Data presented in this paper demonstrate that when the p-value method is used to develop alternate test forms, the resulting tests are likely to meet a statistical test of parallelism and to produce scores that correspond to a degree sufficient for many if not most programmatic purposes.

While these results are suggestive of the strength of the p-value method, the authors are not advocating the necessarily singular use of the method. Clearly, whenever sufficient resources are available, the researcher would be well-advised to use, in tandem, both test- and score-equating methods in order to maximize the information that may be brought to bear on the task of establishing statistical equivalence. There is, in fact, some intuitive appeal in the notion of using a test-equating method followed by application of a score-equating method as a statistical check or "fine tuning" strategy. Should score transformation (equating) prove unnecessary, there would be, at a minimum, the advantages of preserving

the face validity of the percent-correct score and of avoiding the problems of orienting test consumers to the meaning or derivation of transformed scores.

The p-value method has practical virtues beyond consumer receptivity to the scores produced, especially where resources are limited. As against contentions about the limitations of the p-value method (Williams, 1979), experience has shown that extraordinary resources are not required to generate a sufficient item pool for item-by-item p-value substitution, when the developers have expertise in both the subject matter and item construction. Moreover, where two or more parallel forms are administered on an annual basis, the p-value method may provide the most resource-efficient strategy because: (1) it requires smaller sample sizes as compared to the score equating approaches, and (2) the technique of adding "pilot items" to existing forms can be used to generate, during an actual administration, the item statistics (p-values and point biserials) necessary for the item-substitution approach. The latter technique serves as a strategy for reducing the test security problem; it offers the opportunity to eliminate the <u>separate</u> field test of the new test form and the opportunity to develop <u>new</u> parallel forms by substituting new items for some, but not all, of the items on an existing form. In light of both practical considerations and statistical adequacy, the p-value method should be viewed as a viable method for development of parallel test forms--either alone or in combination with other methods.

References

Angoff, W. H. How we calibrate College Board scores. College Board Review, 1968, 68(Summer), 11-14.

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D.C.: American Education, 1971.

Gulliksen, H. Theory of mental tests. New York: John Wiley, 1950.

Nassif, P. M., Pinsky, P. D., & Rubinstein, S. A. Generating parallel test forms for minimum competency exams. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April 1979.

Williams, P. L. Applications of the Rasch model to statewide assessment programs. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

Wright, B. H. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

15
4/80

Further Work Developing Parallel Tests by P-Value Item Substitution

Nassif, Paula M., National Evaluation Systems, Inc.
Pinsky, Paul D., National Evaluation Systems, Inc.
Rubinstein, Sherry A., National Evaluation Systems, Inc.

Generation of parallel test forms, in which statistical and content parallelism is maintained, can be accomplished despite limitations of the assessment setting. Consistency of classification can be ensured across parallel forms while accommodating small numbers of examinees, or frequency of form regeneration by using a p-value item substitution approach.

This paper describes the procedures in generating and fieldtesting "replacement" items, the guidelines in actual item substitution and the verification of parallelism. A third application of this approach is described.